PHÂN TÍCH BAO DỮ LIỆU (DEA) VỚI R

Lê Văn Tuấn

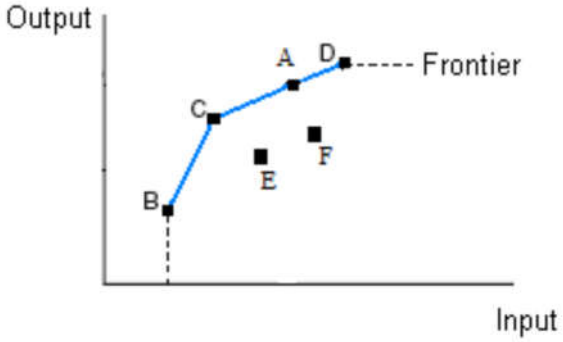
Đại học Thương mại

Tóm tắt. Bài viết giới thiệu về phương pháp phân tích bao dữ liệu (DEA) và cơ sở toán học của phương pháp. Bên cạnh đó, bài viết cung cấp các câu lệnh thực hiện phương pháp DEA trên phần mềm R.
Từ khóa. DEA, phân tích bao dữ liệu, phần mềm R, rDEA

---

1. Giới thiệu phương pháp phân tích bao dữ liệu (DEA)

Phương pháp phân tích bao dữ liệu (Data Envelopment Analysis - DEA) ra đời từ năm 1978, khởi nguồn từ nghiên cứu của Charnes, Cooper và Rhodes, tuy nhiên nó lại có xuất phát điểm từ trước đó hơn 20 năm. Năm 1957, Farrell đưa ra ý tưởng áp dụng đường giới hạn khả năng sản xuất (Production Possibility Frontier – PPF) làm tiêu chí đánh giá hiệu quả (tương đối) giữa các đơn vị (Decision Making Units - DMU, chẳng hạn: công ty, đại lý, trường học,…) trong cùng một ngành; theo đó các đơn vị đạt đến mức giới hạn sẽ được coi là hiệu quả (hơn) và các đơn vị không đạt đến đường PPF sẽ bị coi là kém hiệu quả (hơn các đơn vị kia). Đối với các DMU hiệu quả, vì chúng nằm trên đường giới hạn, nên điểm hiệu quả kỹ thuật (technical efficiency score, gọi tắt là TE) của chúng bằng 1. Đối với các DMU kém hiệu quả (nằm trong đường giới hạn), điểm hiệu quả của chúng sẽ nhỏ hơn 1.



Hình 1. Đường giới hạn khả năng sản xuất

Hiệu quả được tính toán từ đầu ra (outputs) thu được tương ứng với đầu vào (inputs) cho trước. Hình trên minh họa cho trường hợp đơn giản nhất, các đơn vị chỉ có 1 đầu ra và 1 đầu vào. Các đơn vị A, B, C, D là hiệu quả; các đơn vị E, F là không hiệu quả (vì có thể giảm đầu vào nhưng vẫn đạt được đầu ra như trước).

Phương pháp DEA áp dụng bài toán tối ưu hóa tuyến tính phi tham số để xây dựng đường PPF dựa trên số liệu đã biết về một nhóm các đơn vị nhất định và tính toán điểm hiệu quả cho các đơn vị đó. Để minh họa, ta sẽ xét một ví dụ sau.

***Ví dụ*[1].** Một chuỗi cửa hàng bán lẻ có 6 của hàng A, B, C, D, E, F; tại mỗi cửa hàng có 2 đầu vào: số nhân viên và lượng thời gian quản lý (trong tuần), và 2 đầu ra: số quần áo bán được và số phụ tùng bán được (trong tuần). Dữ liệu được cho trong bảng sau.

Bảng 1. Dữ liệu hàng hóa

Cuahang	Nhanvien	Thoigian	Quanao	Phutung
A	51	38	169	119
B	60	45	243	167
C	43	33	173	158
D	53	43	216	138
E	43	38	155	161
F	44	35	169	157

Độ hiệu quả tại mỗi cửa hàng là:

$$Điểm\ hiệu\ quả = \frac{(u_1 * So\ quan\ ao) + (u_2 * So\ phu\ tung)}{(v_1 * So\ nhan\ vien) + (v_2 * Luong\ thoi\ gian)}$$

Với u_1, u_2 và v_1, v_2 tương ứng là trọng số của các đầu ra và đầu vào (có thể xem là giá/ giá thành).

Vì có 6 đơn vị, nên vấn đề trở thành giải 6 bài toán tối ưu, tại mỗi đơn vị cần tối đa hóa hiệu quả với ràng buộc là hiệu quả của các đơn vị khác (với cùng trọng số) không vượt quá 1. Ví dụ với cửa hàng A sẽ là:

Hàm mục tiêu:

$$Maximize \frac{(169 \times u_1) + (119 \times u_2)}{(51 \times v_1) + (38 \times v_2)}$$

Các ràng buộc:

$$\frac{(243 \times u_1) + (167 \times u_2)}{(60 \times v_1) + (45 \times v_2)} \leq 1; \frac{(173 \times u_1) + (158 \times u_2)}{(43 \times v_1) + (33 \times v_2)} \leq 1$$

$$\frac{(216 \times u_1) + (138 \times u_2)}{(53 \times v_1) + (43 \times v_2)} \leq 1; \frac{(155 \times u_1) + (161 \times u_2)}{(43 \times v_1) + (38 \times v_2)} \leq 1$$

$$\frac{(169 \times u_1) + (157 \times u_2)}{(44 \times v_1) + (35 \times v_2)} \leq 1;$$

$$u_1, u_2, v_1, v_2 \geq 0$$

[1] Tham khảo trong [1]

Để đưa về bài toán quy hoạch tuyết tính, ta cố định mẫu số của hàm mục tiêu bằng 1, bài toán trở thành:

Hàm mục tiêu:

$$Maximize(169 \times u_1) + (119 \times u_2)$$

Các ràng buộc:

$$(169 \times u_1) + (119 \times u_2) - (60 \times v_1) + (45 \times v_2) \leq 0$$
$$(173 \times u_1) + (158 \times u_2) - (43 \times v_1) + (33 \times v_2)\ ) \leq 0$$
$$(216 \times u_1) + (138 \times u_2) - (53 \times v_1) + (43 \times v_2) \leq 0$$
$$(155 \times u_1) + (161 \times u_2) - (43 \times v_1) + (38 \times v_2) \leq 0$$
$$(169 \times u_1) + (157 \times u_2) - (44 \times v_1) + (35 \times v_2) \leq 0$$
$$(51 \times v_1) + (38 \times v_2) = 1$$
$$u_1, u_2, v_1, v_2 \geq 0$$

Bài toán tổng quát. Cho N đơn vị, mỗi đơn vị có s đầu vào và m đầu ra. Dữ liệu được cho trong ma trận đầu vào X = (x_{ij}) và đầu ra Y = (y_{ij}). Tìm các bộ trọng số đầu vào (v_1, …, v_s) và đầu ra (u_1, …, u_m) thỏa mãn N bài toán tối ưu. Chẳng hạn tại đơn vị j_0, bài toán sẽ là:

Hàm mục tiêu:

$$Maximize\ \theta_{j_0} = \frac{\sum_{i=1}^{m} u_i y_{i,j_0}}{\sum_{i=1}^{s} v_i x_{i,j_0}}$$

Các ràng buộc:

$$\frac{\sum_{i=1}^{m} u_i y_{i,j}}{\sum_{i=1}^{s} v_i x_{i,j}} \leq 1 \quad \forall j = 1,..,j_0,..N$$

$$u_i \geq 0 \quad \forall i = 1,m;\ v_i \geq 0 \quad \forall i = 1,s$$

Giá trị θ_{j_0} được gọi là điểm hiệu quả (efficiency score).

Bài toán trên có thể đưa về bài toán quy hoạch tuyến tính tương ứng là:

Hàm mục tiêu:

$$Maximize \sum_{i=1}^{m} u_i y_{i,j_0}$$

Các ràng buộc:

$$\sum_{i=1}^{m} u_i y_{i,j} - \sum_{i=1}^{s} v_i x_{i,j} \leq 0 \quad \forall j = 1,..,j_0,..N$$

$$\sum_{i=1}^{s} v_i x_{i,j_0} = 1$$

$$u_i \geq 0 \quad \forall i = 1,m;\ v_i \geq 0 \quad \forall i = 1,s$$

2. Giới thiệu phần mềm R và gói lệnh rDEA

Hiện nay, có khá nhiều phần mềm cho phép ước lượng hiệu quả kỹ thuật theo phương pháp DEA, bao gồm cả phần mềm thương mại (phải mua, ví dụ như DEA Frontier, DEA-Excel-Solver Pro) lẫn phần mềm miễn phí (như DEAP, DEAOS,...). VDEA là tiện ích dùng ngôn ngữ tiếng Việt cho excel được phát triển bởi tác giả Ngô Đăng Thành (xem [3]). Bên cạnh đó, các thư viện miễn phí rDEA, lpSolve, Benchmarking, FEAR của phần mềm R cũng cho phép thực hiện các phân tích tương tự các phần mềm trên.

2.1. Phần mềm R

R là phần mềm (cũng gọi là ngôn ngữ lập trình R) để phân tích dữ liệu được xây dựng bởi Ross Ihaka và Robert Gentleman tại The University of Auckland, New Zealand, tiếp tục được phát triển bởi nhóm R Development Core Team. Phần lớn các kỹ thuật phân tích trong kinh doanh đều được R hỗ trợ: từ Thống kê đến Học máy hay các kỹ thuật Tối ưu hóa. Bằng chứng cho sức mạnh của R đó là những giải thưởng và sự tán dương từ những tạp chí hay cộng đồng uy tín trên thế giới như New York Times, Forbes, Intelligent, Enterprise, InfoWorld và The Register; cũng như được tích hợp phát triển bởi các tập đoàn công nghệ hàng đầu như Microsoft, Google, IBM, Oracle, Amazon-AWS.

Các lí do chính nên sử dụng R trong học thuật cũng như thực tiễn là: Miễn phí (và mã nguồn mở); Phần mềm mạnh nhất trong các phần mềm miễn phí; Cạnh tranh (thậm chí vượt trội) so với các phần mềm thương mại[2]; Đã sử dụng nhiều trong thực tiễn; Chạy được trên nhiều hệ điều hành.

Download và cài đặt trên Windows (R có cả phiên bản trên Linux và (Mac) OS X*)*

- Truy cập vào trang chủ: http://www.r-project.org/, click vào CRAN (dưới chữ Download ở cột bên trái), sẽ đến trang CRAN Mirrors, click vào một link (ví dụ của Thailand), click tiếp Download R for Windows, click tiếp install R for the first time, click tiếp Download R *.*.* for Windows sẽ download được file R-*.*.*-win.exe (*.*.* chỉ version tại thời điểm download).
- Cài đặt như các phần mềm khác.

Thư viện (gói lệnh) của R hỗ trợ phương pháp DEA.

- rDEA

Cài đặt thư viện:

- Tại cửa sổ lệnh của R gõ: install.packages("rDEA")

Sử dụng thư viện:

- Mỗi lần chạy R, tại cửa sổ lệnh gõ: library(rDEA)

2.2. Gói lệnh rDEA

rDEA là gói lệnh/thư viện ứng dụng cho phân tích bao dữ liệu trên R, ưu điểm của nó là khả năng ước tính điểm số DEA một cách mạnh mẽ trong cả hai trường hợp có/không có các biến môi trường và thực hiện các bài kiểm tra tỷ lệ lợi nhuận. Gói lệnh được xây dựng bởi Jaak Simm và Galina Besstremyannaya, phiên bản được công bố vào 6/2/2020 là 1.2-6.

[2] http://stanfordphd.com/Statistical_Software.html

Các chức năng của gói lệnh gồm có (xem [2]):

- dea: Ước tính điểm hiệu quả kỹ thuật (mô hình DEA định hướng đầu vào và đầu ra) và hiệu quả chi phí điểm (DEA tối thiểu hóa chi phí).
- dea.env.robust: Ước tính điểm hiệu quả được điều chỉnh sai lệch trong các mô hình DEA định hướng đầu vào và đầu ra với các biến môi trường (ngoại sinh).
- dea.robust: Thực hiện hiệu chỉnh sai lệch về điểm hiệu quả kỹ thuật của Simar và Wilson (1998) trong đầu vào và các mô hình DEA định hướng đầu ra.
- hospitals: Bộ dữ liệu hành chính cho các biến tài chính và các đặc điểm được lựa chọn của công chúng địa phương tại các bệnh viện ở Nhật Bản.
- multi_glpk_solve_LP: Giao diện R mức cao với Bộ lập trình tuyến tính GNU (GLPK) để giải quyết nhiều bài toán tuyến tính cũng như các bài toán lập trình tuyến tính nguyên hỗn hợp (MILP). Giải quyết nhiều vấn đề cùng lúc và cho phép tránh chi phí giao tiếp R, tối quan trọng khi giải quyết nhiều bài toán nhỏ.
- rts.test: Các kiểm định lợi nhuận trên quy mô trong các mô hình DEA định hướng đầu vào và đầu ra

3. Thực hành ứng dụng

Phần này sẽ trình bày các câu lệnh trên R để minh họa phương pháp DEA giải ***Ví dụ*** ở mục 1.

```
library(rDEA)

data<-read.csv("D:\\data\\dea\\stores.csv", header=TRUE)

data ### Xem dữ liệu

Cuahang Nhanvien Thoigian Quanao Phutung

1       A       51       38     169     119

2       B       60       45     243     167

3       C       43       33     173     158

4       D       53       43     216     138

5       E       43       38     155     161

6       F       44       35     169     157

### Chạy DEA

inp = data[c('Nhanvien', 'Thoigian')]

out = data[c('Quanao', 'Phutung')]
```

```
model = dea(XREF=inp, YREF=out, X=inp, Y=out, model="input", RTS="constant")
model$thetaOpt ### Xem kết quả
[1] 0.8254374 1.0000000 1.0000000 1.0000000 1.0000000 0.9685549
```

Như vậy, ta thấy các cửa hàng hiệu quả là B, C, D, E; các cửa hàng không hiệu quả là A, F.

4. Kết luận

Bài viết đã giới thiệu sơ lược ý tưởng của phương pháp phân tích bao dữ liệu, đồng thời minh họa thực hành ứng dụng trên phần mềm R (với thư viện rDEA) để tính điểm hiệu quả. Có nhiều hướng để có thể nâng cấp chất lượng bài viết, đó là:

- Hiện tại, bài viết mới sử dụng chức năng dea của thư viện rDEA, có thể sử dụng các chức năng khác để áp dụng cho bài toán với biến ngoại sinh.
- Bài viết đang sử dụng bộ số liệu ví dụ đơn giản, có thể áp dụng các tính toán này cho bộ dữ liệu thực tế, từ đó có được những khuyến cáo hữu ích đến các đơn vị, công ty,… nhằm nâng cao hiệu quả sản xuất.

Phương pháp phân tích bao dữ liệu có ý nghĩa cao trong phân tích kinh tế, tuy nhiên, thực chất nó chính là bài toán quy hoạch tuyến tính (được giảng dạy trong học phần Các phương pháp Toán kinh tế của trường Đại học Thương mại). Vì vậy, các tác giả đề xuất giới thiệu phương pháp bao dữ liệu cùng với thực hành trên phần mềm R khi giảng dạy các học phần Kinh tế vi mô, Các phương pháp Toán kinh tế, Các phương pháp phân tích định lượng trong quản lý kinh tế (cao học) tại trường Đại học Thương mại.

TÀI LIỆU THAM KHẢO

1. Bhaskarjit Sarmah. *Introduction to Data Envelopment Analysis in R*. https://medium.com/analytics-vidhya/introduction-to-data-envelopment-analysis-in-r-773745549d6a
2. Jaak Simm và Galina Besstremyannaya. *rDEA: Robust Data Envelopment Analysis (DEA) for R*. https://cran.r-project.org/web/packages/rDEA/rDEA.pdf
3. Ngô Đăng Thành. *Hướng dẫn sử dụng phương pháp Phân tích bao dữ liệu trong Excel: Vietnamese DEA add-in for Excel (phiên bản 2.0).* *https://www.readcube.com/articles/10.2139%2Fssrn.2577136*